Logistics

Class Road Map - IT-I

- L19 (1/6): Overview, Communications, Gaussian Channel
- L20 (1/8): Gaussian Channel, band limitation, parallel channels, optimization and duality
- L21 (1/13): parallel channels, colored noise, feedback, matrix inequalities
- L22 (1/15): matrix inequalities, rate distortion.
  - (1/20): Monday holiday
- L23 (1/22): rate distortion.
- L24 (1/27):
- L25 (1/29):
- L26 (2/3):
- L27 (2/5):
- L28 (2/10):
- L29 (2/12):
- (2/17): Monday, Holiday
- L30 (2/19):
- L31 (2/24):
- L32 (2/26):
- L33 (3/3):
- L34 (3/5):
- L35 (3/10):
- L36 (3/12):

Cumulative Outstanding Reading

- Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
- Read Ch. 9 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book.
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page [http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/]).

Homework

- Homework 1 posted on canvas, due a week from Monday, 1/27/14 at 11:45pm. Only four problems, but these are good problems (and first three are on Gaussian channels so you can start today).
Office hours on Mondays, 3:30-4:30.

As always, email me if you want to skype/google hangout rather than come to office hours, also at different times.

Before and After

Before we had \( k \) independent Gaussian channels with a common power constraint and uncorrelated noise.

Now, have colored noise, i.e.,

\[
Z_{1:k} \sim N(0, K_Z) \tag{22.15}
\]

where \( K_Z \) is not necessarily diagonal (i.e., noise is correlated from one time step to the next).

We still assume \( X_{1:n} \perp \!\!\!\perp Z_{1:n} \) (so noise and signal are independent).
Capacity of Parallel Channels

So the final capacity is then

\[ C_n = \frac{1}{2} \sum_{j=1}^{k} \log(1 + P_i/N_i) \]  \hspace{1cm} (22.15)

\[ = \frac{1}{2} \sum_{j=1}^{k} \log \left( 1 + \frac{(1/\nu - N_i)^+}{N_i} \right) \text{ bits per parallel channel use} \]  \hspace{1cm} (22.16)

In units of bits per transmission (bits per single channel transmission, take the average):

\[ C_n = \frac{1}{2n} \sum_{j=1}^{k} \log \left( 1 + \frac{(1/\nu - N_i)^+}{N_i} \right) \text{ bits per transmission} \]  \hspace{1cm} (22.17)

Colored Noise - KKT

Like before, we have KKT conditions to get:

\[ A_{ii} = (\nu - \lambda_i)^+ \] \hspace{1cm} (22.30)

and

\[ \sum_i (\nu - \lambda_i)^+ = nP \] \hspace{1cm} (22.31)

And the total capacity becomes

\[ C_n = \frac{1}{2n} \sum_{i=1}^{n} \log(1 + \frac{(\nu - \lambda_i)^+}{\lambda_i}) \text{ bits per transmission} \] \hspace{1cm} (22.32)

So this is water filling again, but on the eigenvalues of \( K_Z \).

I.e., we are filling a diagonalized version of the problem. Once we get \( A_{ii} \), we can “re-correlate” using \( K_{X} = QAQ^T \) to get final constraint on \( X \).
Stationary Stochastic (Gaussian) Processes

- So now we have positive semidefinite kernel functions $f_X$ and $f_Z$ such that (stationary case)

$$EX_iX_j = f_X(|i - j|) \quad \text{and } EZ_iZ_j = f_Z(|i - j|) \quad (22.30)$$

so that the matrices are symmetric Toeplitz (i.e., symmetric same along each diagonal)

Spectral Water filling

- Pictured:
Bayesian Network View of Feedback and Colored Noise

With both feedback and correlated noise:

- Note, $Z_i = Y_i - X_i$ which means that
  
  \[ I(X_1:n; Y_1:n) = h(Y_1:n) - h(Y_1:n|X_1:n) \]  
  
  \[ = h(X_1:n + Z_1:n) - h(Z_1:n) \]  

  max when Gaussian  
  Gaussian

- So, max is achieved when $X_{1:n}$ is also Gaussian, and in fact can be jointly Gaussian (exercise: show this).

Capacity with feedback and correlated noise

- Note that $B$ controls the feedback — if $B = 0$ then no feedback.
- This is a general form of conditional Gaussian, where $V$ is also Gaussian and independent of $Z$.
- So we can write this capacity as the following optimization:

  \[ C_{n,FB} = \max \frac{1}{n} \text{tr}(K_X) \leq P \frac{1}{2n} \log \frac{|K_X + Z|}{|K_Z|} \]  

  where $K_X + Z = E[(X + Z)(X + Z)^\top]$  

  \[ = \max_{K_V \succeq 0} \frac{1}{2n} \log \frac{|(B + I)K_Z (B + I)^\top + K_V|}{|K_Z|} \]  

  \[ \text{pred}(B) = \text{true} \]  

  \[ \text{(22.39)} \]

  where $\text{pred}(B) = \text{tr}(BK_Z B^\top + K_V) \leq nP$. Justification on next slide.
Capacity theorem: feedback and correlated noise

Theorem 22.2.2

For a Gaussian channel with feedback, the rate $R_n$ for any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ satisfies $R_n \leq C_{n, FB} + \epsilon_n$ when $\epsilon_n \to 0$ as $n \to \infty$.

Proof.

- Assume $W$ is uniform over $2^{nR}$.
- We have Markov chain

$$W \rightarrow X_{1:n} \rightarrow Y_{1:n} \rightarrow \hat{W} \quad (22.45)$$

- Thus, Fano’s inequality still holds, i.e.

$$H(W|\hat{W}) \leq 1 + nR_nP_e^{(n)} = n\left(\frac{1}{n} + R_nP_e^{(n)}\right) = n\epsilon_n \quad (22.46)$$

where $\epsilon_n \to 0$ as $n \to \infty$ (due to $P_e^{(n)} \to 0$).

---

Matrix Inequalities

Theorem 22.2.2

\(\forall X, Z \ n\text{-D random variables (not necessarily Gaussian), we have}\)

$$K_{X+Z} + K_{X-Z} = 2K_X + 2K_Z \quad (22.53)$$

Proof.

$$K_{X+Z} = E(X + Z)(X + Z)^\top = EXX^\top + EZZ^\top + EZX^\top + EZX^\top = K_X + K_{XZ} + K_{ZX} + K_Z \quad (22.54)$$

$$K_{X-Z} = E(X - Z)(X - Z)^\top = EXX^\top - EZZ^\top - EZX^\top + EZX^\top = K_X - K_{XZ} - K_{ZX} + K_Z \quad (22.55)$$

sum up both to get the result.
Matrix Inequalities

**Theorem 22.2.2**

*If* $A \succeq 0$, $B \succeq 0$, and $A - B \succeq 0$, *then* $|A| \geq |B|$.

**Proof.**

- Let $C = A - B$, $X_1 \sim \mathcal{N}(0, B)$, $X_2 \sim \mathcal{N}(0, C)$, $X_1 \perp X_2$.
- If $Y = X_1 + X_2$ then $Y \sim \mathcal{N}(0, B + C) = \mathcal{N}(0, A)$.
- Then we get
  \[ h(Y) \geq h(Y|X_2) = h(X_1|X_2) = h(X_1) \tag{22.53} \]
- Then, using the entropy of Gaussian formula,
  \[ \frac{1}{2} \log((2\pi e)^n|A|) \geq \frac{1}{2} \log((2\pi e)^n|B|) \tag{22.54} \]
  and the result follows.

---

**Matrix Inequalities**

**Theorem 22.2.2**

*Let* $X, Z$ be arbitrary $n$-D random variables. *Then*

\[ |K_{X+Z}| \leq 2^n |K_X + K_Z| \tag{22.53} \]

**Proof.**

\[
\begin{align*}
2(K_X + K_Z) &= K_{X+Z} + K_{X-Z} \\
\Rightarrow 2(K_X + K_Z) - K_{X+Z} &= K_{X-Z} \succeq 0 \\
\Rightarrow K_{X+Z} &\leq 2(K_X + K_Z) \\
\Rightarrow |K_{X+Z}| &\leq |2(K_X + K_Z)| = 2^n |K_X + K_Z|
\end{align*}
\]
Matrix Inequalities

**Theorem 22.3.1**

Let $A \succeq 0$, $B \succeq 0$, and $0 \leq \lambda \leq 1$, then

$$|\lambda A + (1 - \lambda)B| \geq |A|^\lambda |B|^{1-\lambda}$$  \hspace{1cm} (22.1)

or in other words, $\log(|\lambda A + (1 - \lambda)B|) \geq \lambda \log |A| + (1 - \lambda) \log |B|$ or log-determinant is concave for non-negative definite matrices.

**Proof.**

- Let $X \sim \mathcal{N}(0, A)$, $Y \sim \mathcal{N}(0, B)$
- Let $Z = X\mathbf{1}(\theta = 1) + Y\mathbf{1}(\theta = 0)$, $X \perp \perp Y$, $\theta \perp \perp$ both, and where $P(\theta = 1) = 1 - P(\theta = 0) = \lambda$. Note $Z$ is not necessarily Gaussian.

... proof continued.

- Then
  $$K_Z = EZZ^\top = E[(X\mathbf{1}(\theta = 1) + Y\mathbf{1}(\theta = 0))^{\mathbf{2}}]$$  \hspace{1cm} (22.2)
  $$= E[XX^\top\mathbf{1}(\theta = 1)\mathbf{1}(\theta = 1)] + E[XY^\top\mathbf{1}(\theta = 1)\mathbf{1}(\theta = 0)]$$  \hspace{1cm} (22.3)
  $$+ E[YX^\top\mathbf{1}(\theta = 0)\mathbf{1}(\theta = 1)] + E[YY^\top\mathbf{1}(\theta = 0)\mathbf{1}(\theta = 0)]$$
  $$= \lambda A + (1 - \lambda)B$$  \hspace{1cm} (22.4)

Thus, we get

$$\frac{1}{2} \log((2\pi e)^n|\lambda A + (1 - \lambda)B|) \geq h(Z) \geq h(Z|\theta)$$  \hspace{1cm} (22.5)

$$= \lambda h(X) + (1 - \lambda)h(Y) = \frac{1}{2} \log((2\pi e)^n|A|^\lambda |B|^{1-\lambda})$$  \hspace{1cm} (22.6)

Must we assume $X \perp \perp Y$?
Definition 22.3.2

We define a distribution that factorizes generatively if:

\[ f(x_{1:n}, z_{1:n}) = f(z_{1:n}) \prod_i f(x_i|x_{1:i-1}, z_{1:i-1}) \]  \hspace{1cm} (22.8)

- Note this is called “causally related” in the book, but it need not have anything to do with causality, which can be something entirely different (see the work of J. Pearl for discussions on and models of causality).
- We can view this pictorially:

![Diagram](image1)

Definition 22.3.3

We define a distribution that factorizes generatively if:

\[ f(x_{1:n}, z_{1:n}) = f(z_{1:n}) \prod_i f(x_i|x_{1:i-1}, z_{1:i-1}) \]  \hspace{1cm} (22.9)

- called “causally related” in the book, but it need not have anything to do with causality, which can be something entirely different (see the work of J. Pearl for discussions on and models of causality).
- Pictorially:

![Diagram](image2)
Since feedback codes at capacity are of the form \(X_i = f(Z_{1:i-1}) = f(Z_{1:i-1}, X_{1:i-1})\) (w. implicit \(w\)), the above generative factorization model applies in this case as well.

**Theorem 22.3.4**

*If \(x_{1:n}\) and \(z_{1:n}\) factorize generatively (not nec. Gaussian), then*

\[
h(x_{1:n} - z_{1:n}) \geq h(z_{1:n}) \tag{22.10}
\]

*and*

\[
|K_{X-Z}| \geq |K_Z| \tag{22.11}
\]

**Proof.**

\[
h(X_{1:n} - Z_{1:n}) = \sum_i h(X_i - Z_i | X_{1:i-1} - Z_{1:i-1}) \quad \text{(chain rule)}
\]

\[
\geq \sum_i h(X_i - Z_i | X_{1:i-1}, Z_{1:i-1}, X_i) \quad \text{(conditioning)}
\]

\[
= \sum_i h(Z_i | X_{1:i-1}, Z_{1:i-1}, X_i) \quad \text{(factorization)}
\]

\[
= \sum_i h(Z_i | Z_{1:i-1}) \quad \text{(factorization)}
\]

\[
= h(Z_{1:n}) \quad \text{(chain rule)}
\]
Next, let $\tilde{X}_{1:n}, \tilde{Z}_{1:n}$ be independent Gaussian r.v.'s with covariance $K_X, K_Z$ respectively and let factorization assumption hold. Then

$$\frac{1}{n} \log[(2\pi e)^n |K_{X-Z}|] = h(\tilde{X} - \tilde{Z})$$  \hspace{1cm} (22.17)$$

$$\geq h(\tilde{Z}) = \frac{1}{2} \log((2\pi e)^n |K_Z|)$$  \hspace{1cm} (22.18)$$

and the result follows due to monotonicity of the log.

Now we are ready to start comparing $C_n$ (capacity without feedback) and $C_{n,FB}$ (capacity with feedback).

We'll do this both with an additive bound and a multiplicative bound.

**Additive Bound**

**Theorem 22.3.5**

$$C_{n,FB} \leq C_n + 1/2$$  \hspace{1cm} (22.19)$$

So at most 1/2 bit per channel use of gain!

**Proof.**

$$C_{n,FB} \leq \max_{\frac{1}{n} \operatorname{tr}(K_X) \leq P} \frac{1}{2n} \log \frac{|K_Y|}{|K_Z|}$$  \hspace{1cm} (22.20)$$

$$\leq \max_{\frac{1}{n} \operatorname{tr}(K_X) \leq P} \frac{1}{2n} \log \frac{2^n |K_X + K_Z|}{|K_Z|}$$  \hspace{1cm} (22.21)$$

$$\leq \max_{\frac{1}{n} \operatorname{tr}(K_X) \leq P} \frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|} + \frac{1}{2}$$  \hspace{1cm} (22.22)$$

$$\leq C_n + \frac{1}{2}$$  \hspace{1cm} (22.23)$$
Theorem 22.3.6

\[ C_{n,FB} \leq 2C_n \]  
\[ (22.24) \]

or equivalently

\[ \frac{1}{2n} \log \frac{|KX+Z|}{|KZ|} \leq 2 \frac{1}{2n} \log \frac{|KX+KZ|}{|KZ|} \]  
\[ (22.25) \]

Proof.

\[ \frac{1}{2n} \log \frac{|KX+Z|}{|KZ|} = \frac{1}{2n} \log \frac{|KX+Z|^{1/2}}{|KZ|^{1/2}} \]  
\[ (22.26) \]

\[ = \frac{1}{2n} \log \frac{|KX+Z|^{1/2}|KZ|^{1/2}}{|KZ|} \]  
\[ (22.27) \]

\[ \leq \frac{1}{2n} \log \frac{|KX+Z|^{1/2}|KX-Z|^{1/2}}{|KZ|} \text{ by Thm 22.3.4} \]  
\[ (22.28) \]

\[ \leq \frac{1}{2n} \log \frac{|\frac{1}{2}KX+Z + \frac{1}{2}KX-Z|}{|KZ|} \text{ by Thm 22.3.1} \]  
\[ (22.29) \]

\[ = \frac{1}{2n} \log \frac{|KX+KZ|}{|KZ|} \text{ by Thm 21.6.1} \]  
\[ (22.30) \]
Corollary 22.3.7

\[ C_{n,FB} \leq \min \{2C_n, C_n + 1/2\} \quad (22.32) \]

So unfortunately, feedback in this model is not as useful as we might think it would be.

- We know that the source compresses down to the entropy \( H \), but no further.

- We also know that the signal may be sent through the channel at a rate no more than \( C \).
What if we want to compress $R < H$ or transmit $R > C$? ⇒ Error.
Similarly, what if we allow for errors, but rather than measure error or no error, measure average distortion.
But are all errors created equality? Are all errors as bad as others?
We can measure errors with a distortion function, and we have generalization of the previously stated results.
Rate-distortion curves with achievable region

Vector Quantization

We have symbols $X \in \mathcal{X}$ which could be a continuous or a (say big) discrete domain.
We quantize this region to $\hat{\mathcal{X}}$ where $\hat{\mathcal{X}}$ is discrete and not too big (if $\mathcal{X}$ is discrete, then $|\hat{\mathcal{X}}| \ll |\mathcal{X}|$).

In above, $\hat{\mathcal{X}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_7\}$, $|\hat{\mathcal{X}}| = 7 = M$

There are a set of regions $\mathcal{R}_i$, $i = 1, \ldots, M$, disjoint so that $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for $i \neq j$, and with $\hat{x}_i \in \mathcal{R}_i$ for all $i$. 
Vector Quantization

- The regions cover the entire $\mathcal{X}$ space (i.e., $\bigcup_i \mathcal{R}_i = \mathcal{X}$).
- In general, there are $M$ regions.
- Within each region lies a representative, or exemplar, of that region. That is, for any $x \in \mathcal{R}_i$, then $x$ is represented by, say, $\hat{x}_i \in \mathcal{R}_i$.
- For all $x \in \mathcal{X}$, define function
  $$f : \mathcal{X} \to \{1, 2, \ldots, M\}, \quad \text{with} \quad f(x) = i \text{ where } x \in \mathcal{R}_i \quad (22.33)$$
- For all $x \in \mathcal{X}$, define function
  $$g : \{1, 2, \ldots, M\} \to \mathcal{X}, \quad \text{with} \quad g(i) = \hat{x}_i \quad (22.34)$$
- Thus, $g(f(x)) = \hat{x}_i$ if $x \in \mathcal{R}_i$. I.e., $x$’s representative is $g(f(x))$.
- This is a quantization of $\mathcal{X}$ with codebook the set $\{\hat{x}_i\}_{i=1}^M$.
- Each $\hat{x}_i$ is called a codeword.

Vector Quantization: What determines quality?

- The “quality” of the codebook depends on
  - The specific values $\{\hat{x}_i\}_{i=1}^M$, the representatives.
  - How big $M$ is, or rather $R$, with $M = 2^nR$.
  - How often each of the values within $\{\hat{x}_i\}_{i=1}^M$ are used, or more accurately, a probability distribution $p(x)$ over $\mathcal{X}$.
  - A measure of how bad it is to represent $x \in \mathcal{X}$ by $g(f(x))$, or a distortion $d(x, g(f(x)))$.
  - Expected distortion $D = E_{p(x)}d(X, g(f(X)))$. 
The “Vector” in Vector Quantization

- Now, let’s say that we want to “code” a vector \( x \in \mathcal{X}^n \) (e.g., say an \( n \)-D vector where each element of \( x \) is a member of \( \mathcal{X} \)).
- We quantize the entire space \( \mathcal{X}^n \) into \( M \) disjoint regions \( \{ R_i \} \) that cover \( \mathcal{X}^n \) (so picture before, if square, showed case for \( \mathcal{X}^2 \), or \( n = 2 \)).
- Each vector \( x \in \mathcal{X}^n \) consists of \( n \) source symbols (each source symbol lives in \( \mathcal{X} \)).
- Suppose \( M = 2^{nR} \) for some \( R > 0 \).
- The \( R \) can be viewed as a rate in terms of number of bits per source symbol

\[
\text{rate} = \frac{\log M}{n} = \frac{\log 2^{nR}}{n} = \frac{nR}{n} = R \text{ bits per source symbol} \tag{22.35}
\]

- Intuitively, as rate \( \uparrow \) it should be possible for the distortion \( D \downarrow \)
- Let’s now be a little more formal

Rate distortion: set up

- A source produces \( x_1, x_2, \ldots \sim p(x) \) based on source distribution \( p(x) \) with \( x_i \in \mathcal{X} \) for all \( i \).
- An encoder \( f_n : \mathcal{X}^n \rightarrow \{ 1, 2, \ldots, 2^{nR} \} \) takes a sequence of source symbols \( x_1:n \) and maps them to an integer.
- A decoder \( g_n : \{ 1, 2, \ldots, 2^{nR} \} \rightarrow \hat{\mathcal{X}}^n \) takes an integer and maps to quantized vector (i.e., a codeword).
- A distortion function \( d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+ \) measures how bad the mapping is. I.e., \( d(x, \hat{x}) \) measures the “cost” of representing \( x \in \mathcal{X} \) by \( \hat{x} \in \hat{\mathcal{X}} \).
- Distortion is bounded (sometimes needed) if \( \exists d_{\text{max}} \) such that \( d_{\text{max}} \triangleq \max_{x, \hat{x}} d(x, \hat{x}) < \infty \).
- Ex: Hamming (probability of error) distortion.

\[
d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{otherwise} \end{cases} \tag{22.36}
\]

Then \( Ed(X, \hat{X}) = \Pr(x \neq \hat{x}) \)
Set up

- Another possible distortion for \( x \in \mathbb{R}^n \) might be squared error
  \[ d(x, \hat{x}) = (x - \hat{x})^2, \] or 2-norm
  \[ d(x, \hat{x}) = \|x - \hat{x}\|_2^2 \]
  or \( p \)-norm
  \[ d(x, \hat{x}) = \|x - \hat{x}\|_p^p \]
- We can form the extended distortion as follows
  \[ d(X_1:n, \hat{X}_1:n) \triangleq \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \quad (22.37) \]

\[ \text{(2nR, n) code} \]

**Definition 22.4.1**

A \((2nR, n)\) rate distortion code consists of an encoding function
\[ f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\} \quad (22.38) \]

a decoding function
\[ g_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{\mathcal{X}}^n \quad (22.39) \]

(\text{Note, } H(\hat{\mathcal{X}}^n) \leq nR \text{ since only } 2^{nR} \text{ different codewords.})

The distortion of this code is
\[ D = Ed(X_1:n, g_n(f_n(X_1:n))) = \sum_{x_1:n \in \mathcal{X}^n} p(x_1:n)d(x_1:n, g_n(f_n(x_1:n))) \quad (22.40) \]
Comments

- So very much like compression
- rate of the code $R$ is in bits per source symbol, and bigger rates are worse (from compression ratio point of view).
- Ideally, we want to compress as much as possible while keeping the distortion minimal.
- We have a codebook which consists of $2^{nR}$ codewords.

$$ (g_n(1), g_n(2), \ldots, g_n(2^{nR})) = (\hat{X}^n(1), \hat{X}^n(2), \ldots, \hat{X}^n(2^{nR})) $$

(22.41)

- Each codeword is a vector, i.e., $g_n(i) \in \mathcal{X}^n$.
- The assignment regions are now refereed to using $f_n^{-1}(i)$ for the $i$'th assignment region. Same as $R_i$ from before.
- Due to distortion, quite likely we will have more than one source string map to same target string. I.e., exists $x, x' \in \mathcal{X}$ with $x \neq x'$ such that $g_n(f_n(x)) = g_n(f_n(x'))$.

Achievability and rate-distortion pairs

Definition 22.4.2

A rate-distortion pair $(R, D)$ is said to be achievable if $\exists$ a sequence of $(2^{nR}, n)$ codes $(f_n, g_n)$ with

$$ \lim_{n \to \infty} Ed(X_{1:n}, g_n(f_n(X_{1:n}))) \leq D $$

(22.42)

- So $D$ is the max allowable distortion for code at this rate $R$.
- We can make errors, but not too many (bounded average distortion).
- The type of errors we can make is entirely dependent on the distortion function.
- Def: A rate distortion region for a source is the closure of achievable rate distortion pairs $(R, D)$
- Def: A rate distortion function $R(D)$ is the infimum of rates $R$ such that $(R, D)$ is in rate distortion region. I.e.,

$$ R(D) = \inf \{ R : (R, D) \text{ is achievable} \} $$

(22.43)
Achievability and rate-distortion pairs

• Def: A distortion rate function $D(R)$ is the infimum of distortions $D$ such that $(R, D)$ is in rate distortion region. I.e.,

\[ D(R) = \inf \{ D : (R, D) \text{ is achievable} \} \quad (22.44) \]

• The next definition is very important

Definition 22.4.3

The “information” rate distortion function $R^{(I)}(D)$ for source $X$ and distortion $d(x, \hat{x})$ is defined as

\[ R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \quad (22.45) \]

• Let's now spend a bit of time getting some intuition on this function.

Intuition: Information Rate Distortion Function

For fixed $p(x)$, $I(X; \hat{X})$ is convex in $p(\hat{x}|x)$. ⇒ convex optimization.

We will see in fact how we will use certain methods in convex optimization later (alternating minimization) for computing $R^{(I)}(D)$. 
Intuition: Information Rate Distortion Function

\[ R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{x, \hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \]  

(22.47)

- Related to lossless entropy compression.
- Suppose \( D = Ed(X, \hat{X}) = 0 \) (no distortion) and recall \( d(x, \hat{x}) \geq 0 \) by definition.
- Thus, we must at least have that \( \forall x, \hat{x}: p(x, \hat{x}) > 0, d(x, \hat{x}) = 0 \).
- Consider distortions of the form: \( d(x, \hat{x}) = 0 \Rightarrow \{ x = \hat{x} \} \). For example, \( d(x, \hat{x}) = 1_{\{ x \neq \hat{x} \}} \), or alternatively \( d(x, \hat{x}) = (x - \hat{x})^2 \).
- Thus, \( \forall x, \hat{x} \), if \( p(x, \hat{x}) > 0 \) then \( x = \hat{x} \). Or, the random variables are such that \( X = \hat{X} \).
- Hence,

\[ I(X; \hat{X}) = I(X; X) = H(X) \]  

(22.48)

And if \( p(x) \) is uniform, then

\[ R^{(I)}(D) = H(X) \]  

(22.49)

So, in this case we see that \( D = 0 \) implies \( P_e = 0 \) (zero probability of error).
- In general, does \( \{ P_e = 0 \} \Rightarrow \{ D = 0 \} \)? Yes, as long as \( d(x, x) = 0 \).
- In general, does \( \{ D = 0 \} \Rightarrow \{ P_e = 0 \} \)? No. Consider

\[ d(x, \hat{x}) = \begin{cases} (x - \hat{x})^2 - \tau & \text{if } (x - \hat{x})^2 > \tau \\ 0 & \text{else} \end{cases} \]  

(22.51)
Intuition: Information Rate Distortion Function

\[ R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{x, \hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \quad (22.52) \]

- Intuition, why we want to minimize \( I(X; \hat{X}) \) is that:
  1. It's a process of compression at a particular rate and distortion.
  2. \( I(X; \hat{X}) \) measures the rate of transmission from source symbols to compressed form.
  3. To have a high compression ratio, this rate should be low.
- Consider channel \( X \rightarrow \text{channel} \rightarrow \hat{X} \).
- \( \hat{X} \) used to represent \( X \) (e.g., mp3 file, mp4 video file, etc.)
- Ultimate (lossy) compression would lose all information about \( X \), meaning we’d communicate at a rate of \( R = 0 \leq I(X; \hat{X}) \).
- To compress well, want the communication rate from source \( X \) to its representation \( \hat{X} \) to be small.
- Without distortion constraint, we’d hit \( R = 0 \).
- With distortion constraint \( Ed(X, \hat{X}) \leq D \), this keeps us from communicating at a rate of \( R = 0 \).